In *13th International Conference, Discovery Science*, Vol. 6332
Lecture Notes in Computer Science, B. Pfahringer et al. (Eds.),
pp. 42–56. Springer-Verlag, 2010

An Artificial Experimenter for Enzymatic Response Characterisation

Chris Lovell, Gareth Jones, Steve R. Gunn, and Klaus-Peter Zauner

School of Electronics and Computer Science,
University of Southampton, UK, SO17 1BJ
{cjl07r,gj07r,srg,kpz}@ecs.soton.ac.uk

Abstract. Identifying the characteristics of biological systems through physical experimentation, is restricted by the resources available, which are limited in comparison to the size of the parameter spaces being investigated. New tools are required to assist scientists in the effective characterisation of such behaviours. By combining artificial intelligence techniques for active experiment selection, with a microfluidic experimentation platform that reduces the volumes of reactants required per experiment, a fully autonomous experimentation machine is in development to assist biological response characterisation. Part of this machine, an artificial experimenter, has been designed that automatically proposes hypotheses, then determines experiments to test those hypotheses and explore the parameter space. Using a multiple hypotheses approach that allows for representative models of response behaviours to be produced with few observations, the artificial experimenter has been employed in a laboratory setting, where it selected experiments for a human scientist to perform, to investigate the optical absorbance properties of NADH.

1 Introduction

Biological systems exhibit many complex behaviours, for which there are few models. Take for example the proteins known as enzymes, which are believed to act as biochemical computers [19]. Whilst much is understood within a physiological context, there exists a wide parameter space not yet investigated that may open up the development of biological computers. However, such investigation is restricted by the available resources, which require effective usage to explore the parameter spaces. Biological reactants add an additional problem, as they can undergo undetectable physical changes, which will alter the way they react, leading to observations not representative of the true underlying behaviours. There is therefore need for a new tool, which can aid the creation of response models of biological behaviours. Presented here are artificial intelligence techniques, designed to build models of response behaviours, investigated through an effective exploration of the parameter space. Key to this is the use of a multiple hypotheses technique, which helps manage the uncertainties present in experimentation with few, potentially erroneous, observations. These algorithms, or artificial experimenter, will in the future work with an automated lab-on-chip experiment platform, to provide a fully autonomous experimentation machine.

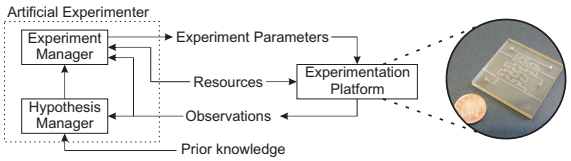


Fig. 1. Flow of experimentation between an artificial experimenter and an automated experimentation platform. A prototype of a lab-on-chip platform in development for conducting the experiments on is shown.

Autonomous experimentation is a closed-loop technique, which computationally builds hypotheses and determines experiments to perform, with chosen experiments being automatically performed by a physical experimentation platform, as shown in Fig.1. Currently few examples of such closed-loop experimentation systems exist in the literature [20, 12, 8], whilst another approach provided the artificial experimenter computational side of the system [9]. Of those that do exist, none consider learning from a small number of observations that could be erroneous. One approach works within a limited domain using extensive prior information to produce a set of hypotheses likely to contain the true hypothesis, allowing the experiment selection strategy to focus on identifying the true hypothesis from a set of hypotheses as cost effectively as possible [8]. However, as such prior knowledge does not exist in the domain of interest to us, our techniques must use also experiments to build a database of information to hypothesise from. Additionally, active learning considers algorithms that sequentially select the observations to learn from [11, 5], however the current literature does not consider learning from small and potentially erroneous sets of observations.

Here we consider the development of an artificial experimenter, where in Section 2 we first consider the issues of building hypotheses in situations where observations are limited and potentially erroneous. Next in Section 3 we consider how such hypotheses can be separated, to efficiently identify the true hypothesis from a set of potential hypotheses, where we introduce a maximum discrepancy algorithm that is able to outperform a selection of existing active learning strategies. In Section 4 we present the design for an artificial experimenter, which is evaluated through simulation in Section 5 and in a laboratory setting in Section 6, to show proof-of-concept of the techniques developed.

2 Hypothesis Manager

The goal for the hypothesis manager is to develop accurate response predictions of the underlying behaviours being investigated, with as few experiments as possible. A key issue is dealing with erroneous observations, which are not representative of the true underlying behaviour being investigated. Whilst the validity of all observations could be determined through repeat experiments, doing so will cut into the resources available for investigating and identifying uncharacterised

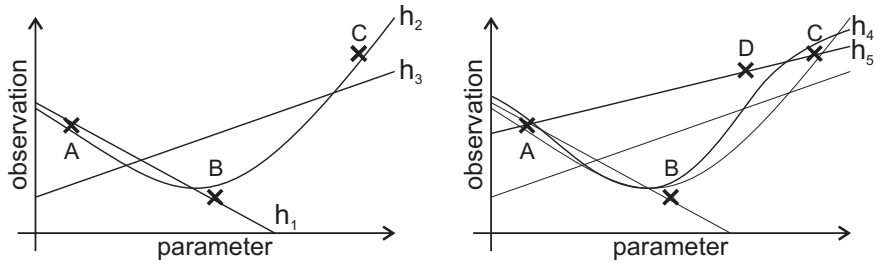


Fig. 2. Validity of observations affecting hypothesis proposal. Hypotheses (lines) are formed after observations (crosses) are obtained. In (a), h_1 formed after A and B are obtained questions the validity of C, whilst h_2 and h_3 consider all observations to be valid with differing levels of accuracy. In (b), D looks to confirm the validity of C, however now h_4 and h_5 differ in opinion about the validity of B.

behaviours. Therefore a hypothesis manager should employ computational methods to handle such uncertainty, built with the view that computation is cheap compared to the cost of experimentation, meaning that computational complexity is unimportant, so long as a solution is feasible.

In experimentation, all observations will be noisy, both in terms of the response value returned and also in the experiment parameter requested. Such noise can be thought of as being Gaussian, until a better noise model can be determined experimentally. As such, we consider a hypothesis as taking the form of a least squares based regression. In particular we use a spline based approach, since it is well defined, can be placed within a Bayesian framework to provide error bars and does not impose a particular spectral scale [18]. A hypothesis is built from a subset of the available observations, a smoothing parameter and a set of weights for the observations, which we will discuss more later.

Erroneous observations however, add a different type of noise, which can be considered as shock noise that provides an observation unrepresentative of the true underlying behaviour. The noise from an erroneous observation is likely to be greater than experimental Gaussian noise, meaning that potentially erroneous observations can be identified as observations that do not agree with the prediction of a hypothesis. The term potentially erroneous is important, as if an observation does not agree with a hypothesis, it may not be the observation that is incorrect, but rather the hypothesis that is failing to model an area of the experiment parameter space. In such limited resource scenarios, when presented with an observation that does not agree with a hypothesis, the hypothesis manager needs to determine whether it is the observation or the hypothesis, or both, which are erroneous.

A possible solution to this problem is to consider multiple hypotheses in parallel, each with a differing view of the observations. Such multiple hypotheses techniques are promoted in philosophy of science literature, as they can ensure

alternate views are not disregarded without proper evaluation, making experimentation more complete [4]. Whilst there are multiple hypotheses based approaches in the literature that produce hypotheses from random subsets of the observations available [6, 1], we believe additional more principled techniques can be applied to aid hypothesis creation. In particular, when a conflicting observation and hypothesis are identified, the hypothesis can be refined into 2 new hypotheses, one that considers the observation to be true, and one that considers the observation to be erroneous. To achieve this, the parameters of the hypothesis are copied into the new hypotheses, however one hypothesis is additionally trained with the potentially erroneous observation having a high weighting, whilst the other is additionally trained with that observation having a zero weighting. By giving the observation a higher weighting, the hypothesis considers the observation to be valid, by having its regression prediction forced closer to that observation. Whilst the zero weighting of the observation makes the hypothesis consider the observation erroneous and removes it from the regression calculation. The handling of potentially erroneous observations through multiple hypotheses, is illustrated in Fig. 2. Next we consider how these hypotheses can be used to guide experiment selection.

3 Effective Separation of the Hypotheses

With the hypothesis manager providing a set of competing hypotheses, there is now the problem of identifying the hypothesis that best represents the true underlying behaviour. To do this we consider methods of separating the hypotheses using experimental design and active learning techniques, evaluated on a simulated set of hypotheses. As the hypotheses will be built from the same small set of observations, their predictions are likely to be similar to each other, with some differences coming from potentially erroneous observations. Therefore, the metric we are interested in, is how well the separation methods perform when the hypotheses have different levels of similarity. To do this the techniques presented will be evaluated using abstract sets of hypotheses, which are described through a single parameter of similarity.

3.1 Techniques

Design of experiments, sequential learning and active learning techniques have considered this problem of hypothesis separation. In particular there is the experimental design technique of T-optimality [2]. However the authors suggest that such designs can perform poorly if the most likely hypothesis is similar to the alternate hypotheses or if there is experimental error [2], which is likely in the experimentation scenario we consider. Whilst many active learning techniques consider this problem in a classification scenario, where there are discrete predictions from the hypotheses [15], meaning that such techniques will require some alteration for a regression problem. The technique we apply to make this alteration, is to use the predictions of the hypotheses as the different classification labels.

In the following, an experiment parameter is represented as x, with its associated observation y. Hypotheses, $h_i(x)$, can provide predictions for experiment parameters through $\hat{h}_i(x)$. Each hypothesis can have its confidence calculated based on the existing observations as:

$$C(h) = \frac{1}{N} \sum_{n=1}^{N} \exp \left(\frac{-\left(\hat{h}(x_n) - y_n\right)^2}{2\sigma^2} \right) \tag{1}$$

where N is the number of observations available. A hypothesis calculates its belief that parameter x brings about observation y through:

$$P_{h_i}(y|x) = \exp \left(\frac{-\left(\hat{h}_i(x) - y\right)^2}{2\sigma_i^2} \right) \tag{2}$$

where σ_i^2 will be kept constant for the abstract hypotheses in the simulated evaluation presented in this section, but is substituted for the error bar of the hypothesis when applied to real hypotheses discussed in Section 5 and Section 6. Additionally, where observations are to be predicted, the hypotheses provide predictions through substituting y for $\hat{h}(x)$. Finally, the working set of hypotheses under consideration is defined as $\mathcal{H}$, which has a size of $|\mathcal{H}|$. We now consider different active learning techniques.

Variance The difference amongst a group of hypotheses has been previously considered through looking at the variance of the hypotheses predictions [3]. Experiments are selected where the variance of the predictions is greatest. So as to allow for previous experiments to be taken into consideration on subsequent calls to the experiment selection method, the confidence of the hypothesis can be used to provide a weighted variance of the predictions, based on how well each hypothesis currently matches the available observations:

$$x^*_{\text{Var}} = \arg\max_x \; k \sum_{i=1}^{|\mathcal{H}|} C\left(h_i\right) \left(\hat{h}_i(x) - \mu^*\right)^2 \tag{3}$$

where

$$\mu^* = \frac{1}{\sum_{i=1} C(h_i)} \sum_{i=1}^{|\mathcal{H}|} C(h_i) \hat{h}_i(x) \tag{4}$$

and k is a normalising constant for weighted variance.

KL Divergence The Kullback-Liebler divergence [10], has been employed as a method for separating hypotheses where there are discrete known labels [13]:

$$x^*_{\text{KLM}} = \arg\max_x \frac{1}{|\mathcal{H}|} \sum_{i=1}^{|\mathcal{H}|} \sum_{j=1}^{|\mathcal{H}|} P_{h_i}\left(\hat{h}_j(x)|x\right) \log \frac{P_{h_i}\left(\hat{h}_j(x)|x\right)}{P_{\mathcal{H}}\left(\hat{h}_j(x)|x\right)} \tag{5}$$

where

$$P_{\mathcal{H}}\left(\hat{h}_j(x)|x\right) = \frac{1}{|\mathcal{H}|}\sum_{k=1}^{|\mathcal{H}|} P_{h_k}\left(\hat{h}_j(x)|x\right) \tag{6}$$

which is the consensus probability between all hypotheses that the observation y_j will be obtained, within some margin of error, when experiment x is performed. This discrepancy measure selects the experiment that causes the largest mean difference between the individual hypotheses and the consensus over the observation distributions.

In its current form this approach requires hypotheses that do not match the observations to be removed. However, if $P_{h_i}(\hat{h}_j(x)|x)$ is multiplied by the confidence of the hypothesis, $C(h_i)$, and the normalising term $\frac{1}{|\mathcal{H}|}$ in (5) and (6) is replaced with the inverse of sum of the confidences, $\frac{1}{C}$, the impact a hypothesis has on the decision process can be scaled by its confidence.

Bayesian Surprise The KL divergence has also been applied to formulate a notion of surprise, within a Bayesian framework [7]. The prior probability is determined from the available observations:

$$P_{h_i}(Y|X) = \frac{1}{n}\sum_{j=1}^{n} P_{h_i}(y_j|x_j) \tag{7}$$

Whilst the predicted posterior probability also takes into consideration what the new probability of the hypothesis would be if a particular experiment x_p was performed that resulted in a specific y_p:

$$P_{h_i}(Y, y_p|X, x_p) = \frac{1}{n+1}\left(nP_{h_i}(Y|X) + P_{h_i}(y_p|x_p)\right) \tag{8}$$

Using these distributions, we consider all predicted observations to determine a surprise term:

$$x^*_{\text{surprise}} = \arg\min_x \frac{1}{|\mathcal{H}|}\sum_{i=1}^{|\mathcal{H}|}\sum_{j=1}^{|\mathcal{H}|} K\left(h_i, \hat{h}_j(x)\right) \tag{9}$$

where K is the KL divergence to provide Bayesian surprise [7]

$$K\left(h_i, y_j\right) = P_{h_i}\left(Y, y_j|X, x\right)\log\frac{P_{h_i}\left(Y, y_j|X, x\right)}{P_{h_i}\left(Y|X\right)} \tag{10}$$

Importantly the experiment with the lowest KL divergence is selected, so as to find the experiment that weakens all hypotheses. If the maximum value were used, it would select the experiment that improves all hypotheses, which by definition will limit the difference between the hypotheses. It can be shown using the framework presented here, that using the minimum KL divergence value results in a better performing discrepancy technique than using the maximum KL divergence.

Maximum Discrepancy Separating the hypotheses can be thought of as identifying experiments that maximise the disagreement between the predictions of hypotheses. Mathematically we consider maximising the integration of the differences between all of the hypotheses, over all possible experiment outcomes:

$$A = \sum_{i=1}^{|\mathcal{H}|} \sum_{j=1}^{|\mathcal{H}|} \int (h_i - h_j)^2 \, dy_t \tag{11}$$

where the likelihood function $P_h(y|x)$ can be used to determine the differences in the hypotheses:

$$A = \sum_{i=1}^{|\mathcal{H}|} \sum_{j=1}^{|\mathcal{H}|} \int \left(P_{h_i}(y|x) - P_{h_j}(y|x)\right)^2 dy \tag{12}$$

then as $P_{h_i}(y|x)$ is a Gaussian distribution, and distinct y can be taken from the predictions of the hypotheses, we can formulate a discrepancy measure:

$$x^*_{\text{discrepancy}} = \arg\max_x \sum_{i=1}^{|\mathcal{H}|} \sum_{i=j}^{|\mathcal{H}|} 1 - P_{h_i}\left(\hat{h}_j(x)|x\right) \tag{13}$$

where we look for the experiment parameter where the hypotheses disagree the most. Next a method of using the prior information is required. On subsequent runs, the discrepancy within the sets of currently agreeing hypotheses should be found, whilst also taking into consideration how well those hypotheses fit the observations. The disagreement term, $1 - P_{h_i}(y_j|x)$, can therefore be multiplied by $P(h_i, h_j|\mathcal{D})$, defined as:

$$P(h_i, h_j|\mathcal{D}) = C(h_i)C(h_j)S(h_i|h_j) \tag{14}$$

where

$$S(h_i, h_j) = \frac{1}{N} \sum_{n=1}^{N} \exp\left(\frac{-\left(\hat{h}_i(x_n) - \hat{h}_j(x_n)\right)^2}{2\sigma_i^2}\right) \tag{15}$$

is the similarity between two hypotheses predictions for the previously performed experiments, with σ_i coming from the error bar of h_i at x for real hypotheses, and is kept constant in the abstract trial discussed next.

3.2 Hypothesis Separation Results

To evaluate the experiment selection techniques, an arbitrary function is used to create a set of potential training observations. These observations are distorted from the function through Gaussian noise, where the amount of noise is the parameter that controls how different the hypotheses in the set are. Twenty hypotheses are then trained from random subsets of the training observations using an arbitrary regression technique. The hypotheses are then compared to

Table 1. Number of experiments until the hypothesis with the highest confidence is the true hypothesis. The similarity is shown as the Gaussian noise applied to the initial training data, where a noisier set of training data provides hypotheses less similar to each other. The best strategy in each case is highlighted in bold.

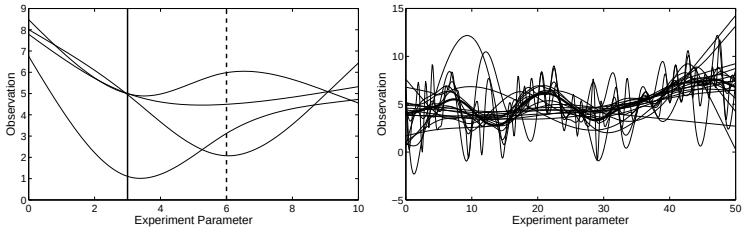
Hypothesis Similarity (increasing order)	Strategy				
	Random	Variance	Max Discrepancy	Surprise	KL Divergence
$N(0, 4^2)$	3	**2**	**2**	3	**2**
$N(0, 2^2)$	8	4	**3**	7	4
$N(0, 1^2)$	18	**7**	**7**	13	11

each other, with the hypothesis that is most similar to all other hypotheses being chosen to act as the true hypothesis. The training observations are then discarded. The true hypothesis provides the observations for the experiments that the active learning techniques request, distorted by Gaussian noise $N(0, 0.5^2)$. The goal is for the active learning techniques to provide evidence to make the true hypothesis have the sustained highest confidence of all the hypotheses in consideration, where the techniques do not know which is the true hypothesis.

Shown in Table 1 are the results for the average number of experiments, over 100 trials, required for the most confident hypothesis to be the true hypothesis, for sets of hypotheses of increasing similarity. As the similarity between the hypotheses increases, it is clear that the variance and maximum discrepancy experiment selection techniques provide the most efficient methods for selecting experiments to separate the hypotheses. However, the variance approach can suffer if there is a hypothesis that makes a prediction that is significantly different to the other hypotheses. As illustrated in Fig. 3(a), alongside an example set of hypotheses in (b), the variance approach can select an experiment where the majority of the hypotheses have the same view, which will likely result in no information gain from that observation. The maximum discrepancy approach however, provides a more robust approach at selecting experiments to separate hypotheses and as such, it will form the basis for the experiment selection strategy employed by the artificial experimenter. The design of which we discuss in the next section.

4 Artificial Experimenter

Building on the concepts discussed earlier of multiple hypotheses and maximum discrepancy experiment selection, we now discuss the design of the artificial experimenter. To begin a number of exploratory experiments are performed, positioned equidistant in the parameter space. In the simulated and laboratory evaluation, 5 experiments are initially performed. After these experiments are performed, an initial set of working hypotheses are created using random subsets of the available observations and randomly selected smoothing parameters. The smoothing parameter is chosen from a set of predetermined smoothing parameters that allow for a range of fits. Initially 200 hypotheses are created in this



(a) (b)

Fig. 3. In (a) is an illustration of where the variance approach can fail, where the solid line is the experiment parameter chosen by the variance approach and dashed is where the maximum discrepancy approach chooses, for the hypotheses shown as curved lines. The variance approach is mislead by a single hypothesis. In (b) is an example set of hypotheses used to test the different active learning techniques for separating a corpus of similar hypotheses, where the bold hypothesis is the true hypothesis.

manner. The observations are then compared against all of the hypotheses to find observations that do not agree with the hypotheses. An observation is determined to be in disagreement with a hypothesis, if that observation is outside the 95% error bar for the hypothesis. If a hypothesis and observation disagree, the parameters of the hypothesis are used to build 2 new hypotheses. These 2 new hypotheses are refinements, where one hypothesis will consider the observation as valid by applying a weight of 100 to the observation, whilst the other hypothesis considers the observation erroneous by applying a weight of 0 to the observation. All 3 hypotheses are then retained in the working set of hypotheses.

After this process of refinement, the hypotheses are evaluated against all available observations, using the confidence function in Eqn. 1. For computational efficiency, the worst performing hypotheses can at this stage be removed from the working set of hypothesis. Currently the best 20% hypotheses are kept into the next stage of experimentation, as initial tests have indicated that higher percentages provided little additional benefit and only increased the computational complexity.

Next a set of experiments to perform are determined by evaluating the hypotheses with the discrepancy equation:

$$D(x) = \sum_{i=1}^{|\mathcal{H}|} \sum_{j=1}^{|\mathcal{H}|} \left(1 - P_{h_i}\left(\hat{h}_j(x)|x\right)\right) C(h_i)C(h_j)S(h_i,h_j) \tag{16}$$

with the error bars of the hypotheses providing σ_i for Eqn. 2. Whilst the discrepancy approach has been shown to be efficient in identifying the best fitting hypothesis from a set of hypotheses, it is not designed to explore the parameter space to help build those hypotheses. Therefore, to promote exploration, the peaks of Eqn. 16 are used to determine the locations for the set of experiments to next perform. This allows experiments to be performed that investigate differences between the hypotheses in several areas of the parameter space. Addi-

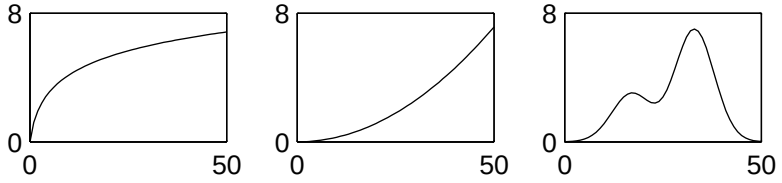


Fig. 4. Underlying behaviours used to evaluate the artificial experimenter, motivated from possible enzyme experiment responses.

tionally, repeat experiments are not performed. The set of experiments are then performed sequentially, where after each experiment is performed, a new set of hypotheses are created, merged with the working hypotheses, which are refined, evaluated and reduced in the process described previously. Once all experiments in the set are performed, a new set of experiments are determined by evaluating the current working set of hypotheses with the discrepancy equation again.

5 Simulated Results

Evaluating the ability of the technique to build suitable models of biological response characteristics, requires underlying behaviours to compare the predictions against. Whilst documented models of the enzymatic behaviours to be investigated do not exist, there are some possible characteristics that may be observed defined in the literature. In Fig. 4 we consider three potential behaviours, motivated from the literature, where (a) is similar to Michaelis-Menton kinetics [14], (b) is similar to responses where there is a presence of cooperativity between substrates and enzymes [17], whilst (c) considers nonmonotonic behaviours that may exist in enzymatic responses [19].

To perform the simulation, we assume that a behaviour being investigated is captured by some function $f(x)$. Calls to this function produce an observation y, however, experimental noise in both the observations obtained (ϵ) and the experiment parameters (δ), deviate this observation from the true response. Additionally, erroneous observations can in some experiments occur through a form of shock noise (ϕ). Whilst ϵ and δ may occur in all experiments, represented through a Gaussian noise function, ϕ will only occur for a small proportion of experiments and will be in the form of a larger offset from the true observation. Therefore we use the following function to represent performing an experiment:

$$y = f(x + \delta) + \epsilon + \phi \tag{17}$$

with the goal of the artificial experimenter being to determine a function $g(x)$ that suitably represents the behaviours exhibited by $f(x)$.

In the simulation, $\epsilon = N(0, 0.5^2)$ for all experiments and $\phi = N(3, 1)$ for 20% of the experiments performed, with one of the first 5 being guaranteed to

be erroneous. Shock noise δ is currently not used for clarity of results. In each trial, 5 initial experiments are performed, with a further 15 experiments being chosen through an active learning technique. In addition to the multiple hypotheses approach presented here, for comparison a single hypothesis approach is tested that is trained with all available observations, using cross-validation to determine the smoothing parameter. The single hypothesis approach is evaluated using two experiment selection methods, which are random selection and placing experiments where the error bar of the hypothesis is maximal. The multiple hypotheses approach is evaluated using three experiment selection methods, which are random selection, the multiple peaks of the discrepancy equation as presented previously, and choosing the single highest peak of the discrepancy equation for each experiment. For each technique and underlying behaviour, 100 trials are conducted, with the bias and variance of the most confident hypothesis of each trial compared to the true underlying behaviour, being used to evaluate the techniques:

$$E = \frac{1}{N}\sum_{n=1}^{N}\left(\left(\bar{b}(x_n) - f(x_n)\right)^2 + \frac{1}{M}\sum_{m=1}^{M}\left(\hat{b}_m(x_n) - \bar{b}(x_n)\right)^2\right) \tag{18}$$

where $\bar{b}(x_n)$ is the mean of the predictions of the most confident hypotheses, $\hat{b}_m(x_n)$ is the prediction of the most confident hypothesis in trial m, M is the number of trials and N is the number of possible experiment parameters.

In Fig. 5, the performance of the different artificial experimenter techniques are shown. The single hypothesis approaches only perform well in the monotonic behaviours shown in (a) and (b), as the cross-validation allows for errors to be smoothed out quickly. However, in the nonmonotonic behaviour, the single hypothesis approaches perform worse, as the features of the behaviours are smoothed out by the cross-validation, as shown in (d), where the single hypothesis approach misses the majority of the features in the behaviour. On the other hand, the multiple hypotheses approach using the presented technique, fairs well in all behaviours. After 15 experiments it has the lowest prediction error of the techniques tested all three behaviours tested here. However, the multiple hypotheses approach using random experiment selection, is able to reduce the error at a faster rate in the nonmonotonic behaviour (c). Whilst as expected, choosing the single highest peak in the discrepancy equation after each experiment, performs the worst of the multiple hypotheses techniques as expected throughout, as that approach does not effectively explore the parameter space.

The difference between the random and multiple peaks experiment selection strategy, is due to the multiple peaks strategy initially finding the differences between hypotheses that poorly represent the underlying behaviour. These early experiments will investigate discrepancies that will return more general information about the behaviour, with it being possible for experiments within a particular set obtaining similar information. However, as the hypotheses better represent the underlying behaviour, the discrepancies between the hypotheses are more likely to indicate where more specific differences in the hypotheses exist, for example a smaller peak in the behaviour being investigated. This is why

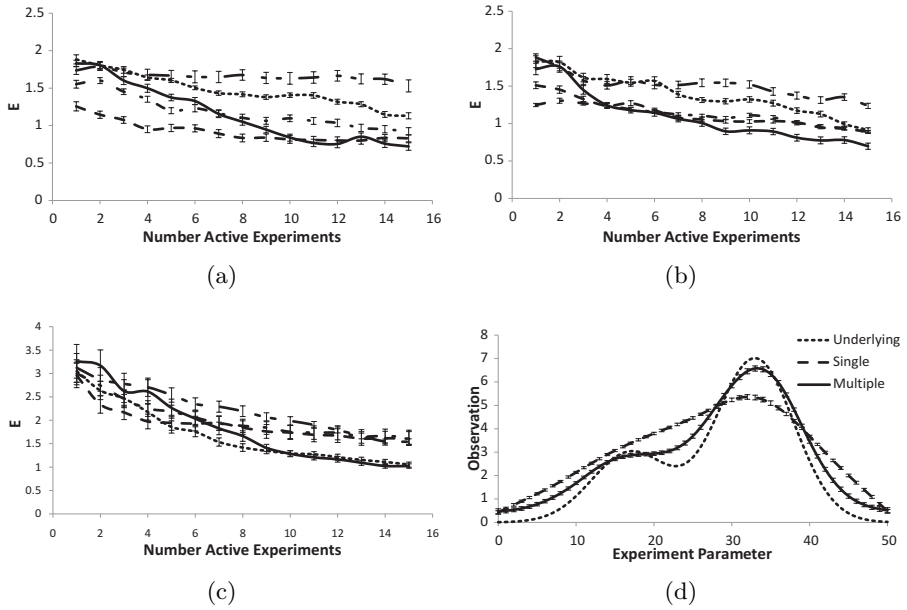
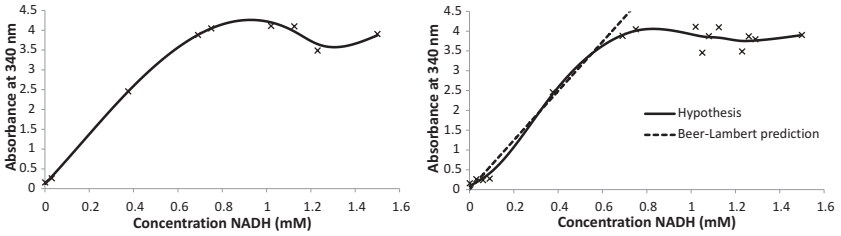


Fig. 5. Comparison of error over number of actively chosen experiments, where 20% of the observations are erroneous in (a-c), with comparison to true underlying for (c) shown in (d). Figures (a-c) correspond to the behaviours in Fig. 4. In (a-c) the lines represent: single hypothesis - variance (dashed), single hypothesis - random (dash dot), multiple hypotheses - discrepancy peaks (solid), multiple hypotheses - random (dots), multiple hypotheses - single max discrepancy (dash dot dot). The multiple hypotheses technique using the peaks of the discrepancy function provides the lowest error after 15 actively selected experiments consistently. The single hypothesis approach fails to identify features in nonmonotonic behaviours shown in (d).

in all three of the behaviours tested, the multiple peaks experiment selection strategy is initially one of the worst performing strategies, but then reduces its error at a faster rate than any of the other strategies. These results suggest that the multiple peaks experiment strategy may in some scenarios benefit from additional exploration, before the active strategy begins. Next we consider an evaluation of the technique within a laboratory setting.

6 Laboratory Evaluation

Further to the simulated evaluation, the artificial experimenter has been tested within a real laboratory setting. Here the artificial experimenter has guided a human scientist to characterise the optical absorbance profile of the coenzyme NADH, where the rate of change of absorbance can be compared to the Beer-Lambert law. NADH is commonly used for monitoring enzymatic catalytic activity.

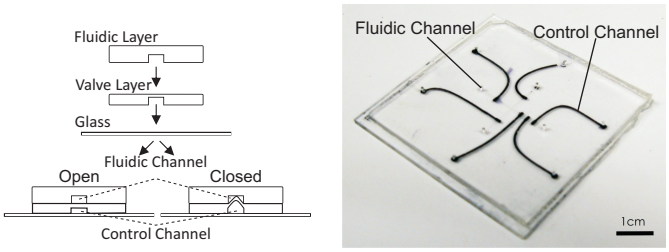


(a) (b)

Fig. 6. Most confident hypothesis and experiments chosen for NADH absorbance characterisation. A stock solution of 5 mM NADH and a 10 mM Tris buffer at pH 8.5 were prepared. Dilutions of NADH requested by the artificial experimenter were produced by mixing volumes taken from the stock solution and the buffer. Measurements of optical absorbance at 340 nm were recorded with a PerkinElmer Lambda 650 UV-Vis Spectrophotometer to provide the observations. The photometric range of the spectrophotometer was 6 A. In (a) the most confident hypothesis after 4 active experiments is shown, where a slight dip in absorbance has been detected. Further experiments determine this dip does not exist, as shown in (b). The hypothesis identifies a linear region in good agreement with the Beer-Lambert law, whilst also identifying a nonlinear region that is likely caused by nonlinear optical effects, as all measurements are within the operational range of the spectrophotometer.

To perform the test, the artificial experimenter was first provided the boundary to which it could explore in the parameter space, 0.001–1.5 mM. The parameter space was coded to the parameter space used in the simulation, allowing for same set of smoothing parameters to be used ($\lambda = \{10, 50, 150, 100, 500, 1000\}$). The artificial experimenter requested an initial 5 experiments, placed equidistant within the parameter space. Using the procedure described in Fig.6, the human scientist performed the experiments as directed, providing the observations to the artificial experimenter. The artificial experimenter then presented a graph of the observations, along with the current best hypothesis, the alternate hypotheses and the discrepancy amongst them. The artificial experimenter was then allowed to select an additional 10 experiments using the multiple peaks active experiment selection technique described.

In Fig 6, the results of those experiments are shown. After the initial exploratory experiments, the artificial experimenter identifies the key feature that there is an increase in absorbance between 0.001 and 0.75 mM, that then begins to level off. The first active experiment looks at roughly where the increase in absorbance ends at 0.69 mM, with the observation agreeing with the initial trend of the data. The second active experiment at 1.23 mM, providing an observation lower than the initial prediction, makes the artificial experimenter consider the possibility that rather than a leveling off in absorbance, the absorbance lowers again with a similar rate to that which it increased. The remainder of experiments then look to investigate whether the absorbance lowers or remains largely flat, with a few additional experiments investigating where the rise in absorbance



(a) (b)

Fig. 7. Microfluidic chip layered design (left) and photo of prototype chip (right). Reactants flow in channels between the fluidic and valve layers, whilst control channels exist between the valve and glass layers. Pressure on the control channels control whether fluidic channels are open or closed, to allow reactants to pass. On-chip absorbance measurement will allow for all experimentation to take place on chip.

begins. The hypothesis after 15 experiments matches the expected Beer-Lambert law rate of change in absorbance prediction, using the indicated extinction coefficient of 6.22 at a wavelength of 340 nm [16], as shown in Fig. 6(b).

7 Conclusion

Presented here is an artificial experimenter that can direct experimentation in order to efficiently build response models of behaviours, where the number of experiments possible is limited and the observations are potentially erroneous. The domain of enzymatic experiments is used to motivate the approach, however the technique is designed to be general purpose and could be applied to other experimentation settings where there are similar limiting factors. The technique uses a multiple hypotheses approach, where different views of the observations are taken simultaneously, in order to deal with the uncertainty that comes from having potentially erroneous observations and limited resources to test them. A technique of experiment selection that places experiments in locations of the parameter space where the hypotheses disagree has been proposed. Whilst this approach appears to perform consistently across simulated behaviours, perhaps additional measures of exploration could be added to the technique, so as to better manage the exploration-exploitation trade-off. Additionally the approach should also consider when to terminate experimentation by monitoring the change in hypotheses over time, rather than using fixed numbers of experiments allowed.

The next stage is to couple the artificial experimenter with the lab-on-chip experiment platform in development, which is shown in Fig. 7. This autonomous experimentation machine, will allow the artificial experimenter to request experiments to be performed, which the hardware will automatically perform, returning the result of the experiment back to the computational system. As such, it will provide a tool for scientists, which will not only allow them to reduce experimentation costs, but will also allow them to redirect their time from monotonous

characterisation experiments, to analysing the results, building theories and determining uses for those results.

Acknowledgements The reported work was supported in part by a Microsoft Research Faculty Fellowship to KPZ.

References

1. Abe, N., Mamitsuka, H.: Query learning strategies using boosting and bagging. In: ICML '98. pp. 1–9. Morgan Kauffmann, San Francisco, CA, USA (1998)
2. Atkinson, A.C., Fedorov, V.V.: The design of experiments for discriminating between several models. Biometrika 62(2), 289–303 (1975)
3. Burbidge, R., Rowland, J.J., King, R.D.: Active learning for regression based on query by committee. In: IDEAL 2007. pp. 209–218. Springer-Verlag (2007)
4. Chamberlin, T.C.: The method of multiple working hypotheses. Science (old series) 15, 92–96 (1890), reprinted in: Science, v. 148, p. 754–759, May 1965.
5. Cohn, D.A., Ghahramani, Z., Jordan, M.I.: Active learning with statistical models. Journal of Artificial Intelligence Research 4, 129–145 (1996)
6. Freund, Y., Seung, H.S., Shamir, E., Tishby, N.: Selective sampling using the query by committee algorithm. Machine Learning 28, 133–168 (1997)
7. Itti, L., Baldi, P.: Bayesian surprise attracts human attention. Vision Research 49, 1295–1306 (2009)
8. King, R.D., Whelan, K.E., Jones, F.M., Reiser, P.G.K., Bryant, C.H., Muggleton, S.H., Kell, D.B., Oliver, S.G.: Functional genomic hypothesis generation and experimentation by a robot scientist. Nature 427, 247–252 (2004)
9. Kulkarni, D., Simon, H.A.: Experimentation in machine discovery. In: Shrager, J., Langley, P. (eds.) Computational Models of Scientific Discovery and Theory Formation, pp. 255–273. Morgan Kaufmann Publishers, San Mateo, CA (1990)
10. Kullback, S., Leibler, R.A.: On information and sufficiency. Annals of Mathematical Statistics 22, 79–86 (1951)
11. MacKay, D.J.C.: Information–based objective functions for active data selection. Neural Computation 4, 589–603 (1992)
12. Matsumaru, N., Colombano, S., Zauner, K.-P.: Scouting enzyme behavior. In: CEC. pp. 19–24. IEEE, Piscataway, NJ, Honolulu, Hawaii (2002)
13. McCallum, A.K., Nigam, K.: Employing em and pool-based active learning for text classification. In: ICML. pp. 584–591. Morgan Kaufmann (1998)
14. Nelson, D.L., Cox, M.M.: Lehninger Principles of Biochemistry. W. H. Freeman and Company, New York, USA, 5th edn. (2008)
15. Settles, B.: Active learning literature survey. Tech. rep., University of Wisconsin-Madison (2009)
16. Siegel, J.M., Montgomery, G.A., Bock, R.M.: Ultraviolet absroption spectra of dpn and analogs of dpn. Archives of Biochemistry and Biophysics 82(2), 288–299 (1959)
17. Tipton, K.F.: Enzyme Assays, chap. 1, pp. 1–44. Practical Approach, Oxford University Press, Oxford, England, 2nd edn. (2002)
18. Wahba, G.: Spline Models for Observational Data, CBMS-NSF Regional Conference series in applied mathematics, vol. 59. SIAM, Philadelphia, PA (1990)
19. Zauner, K.-P., Conrad, M.: Enzymatic computing. Biotechnol. Prog. 17, 553–559 (2001)
20. Żytkow, J., Zhu, M., A.Hussam: Automated discovery in a chemistry laboratory. In: AAAI-90. pp. 889–894. AAAI Press / MIT Press (1990)